Analisi Predittiva

CT0429
Quarto appello

Agosto, 2022

Cognome: _____ Nome: _____

Matricola: _____ Firma: _____

ISTRUZIONI
(DA LEGGERE ATTENTAMENTE).

Assicuratevi di aver scritto nome cognome e matricola sia qui che sul file Rmarkdown disponibile su Moodle. Il tempo a disposizione per completare tutto l'esame (la parte scritta e la parte su Moodle) è di **90 minuti**.

Nessuno studente può lasciare l'aula fino a che la docente non avrà verificato che tutti abbiano consegnato sia il compito scritto che il file Rmarkdown. Dopo la consegna attendete che la docente dia il permesso di lasciare l'aula.

Question 1 (4 points)

Un app di smart-mobility ha lanciato una campagna promozionale in cui agli utenti già iscritti che invitano nuovi utenti viene riconosciuto un credito di 10 euro. In prima istanza, si costruisce un modello per indagare i fattori che influenzano la scelta dell'utente di inviare almeno un invito a nuovi utenti. Le variabili a disposizione sono le seguenti:

- `Invite`: una variabile dicotomica che indica se l'utente ha inviato almeno un invito.
- `Usage`: una variabile categoriale che indica quanto frequentemente l'utente usa la app. La variabile prende valori `Rare`, `Medium`, `Frequent`.
- `Distance`: la distanza media mensile (in centinaia Km) per cui l'utente ha utilizzato i servizi della app.
- `TimeRegister`: i giorni trascorsi dalla data in cui si è registrato l'utente

```
(Intercept)  UsageMedium    UsageRare     Distance TimeRegister
    -0.4120      -0.4645      -0.6715       1.2695       0.0012
```

Viene stimato un modello lineare generalizzato con variabile risposta binomiale:

```
fit <- glm(Invite ~ Usage+Distance+TimeRegister, data = df, family = binomial)
coef(fit)
```

```
(Intercept)  UsageMedium    UsageRare     Distance TimeRegister
    -0.4120      -0.4645      -0.6715       1.2695       0.0012
```

Viene stimato il seguente modello:

(i) Si interpreti il valore del coefficiente angolare relativo alla variabile `Distance`

Dal momento che si stima un GLM della famiglia binomiale, si assume che la funzione legame utilizzata sia la funzione canonica (funzione logistica). Il coefficiente angolare relativo alla variabile **Distance** rappresenta l'effetto sulla scala della funzione logistica del predittore sull'odds ratio.

Tra due istanze (assumendo che gli altri predittori abbiano valori fissati), al crescere di un'unita' di misura di Distance (espressa come distanza media mensile in centinaia di km), in media vi e' un aumento della probabilita' che l'utente abbia inviato almeno un invito (Y=1), evidenziato dalla relazione positiva. Questo aumento pero' non e' costante, poiche' mediato dalla funzione logit.

(ii) Per quale di questi due utenti è più alta la probabilità che venga mandato almeno un invito ad un nuovo utente?

```
d
```

	Usage	Distance	TimeRegister
Utente 1	Rare	80	2
Utente 2	Rare	80	10

E[Y=1| XBeta] = exp(t)/(1+exp(t))

t1 = -1.0835 + (1.2695*0.8) + (0.0012*2) = -0.0655
t2 = -1.0835 + (1.2695*0.8) + (0.0012*10) = -0.0559

E[Y=1| t1] = 0.4836
E[Y=1| t2] = 0.4860

(iii) Quale categoria di utenti è più probabile invii almeno un invito ad un nuovo utente?

Il modello, includendo i fattori, calcola 3 intercette.

t1: UsageFrequent -> Intercetta = -0.4120
t2: UsageMedium -> Intercetta = -0.8855
t3: UsageRare -> Intercetta = -1.0835

A parita' di valori (TimeRegister = 10, Distance = 80)

E[Y=1|t1] = 0.6471648
E[Y=1|t2] = 0.5354653
E[Y=1|t3] = 0.4860286

Question 2 (7 points)

Un coltivatore d'uva studia l'effetto di svariate variabili sulla resa di uva in diversi appezzamenti di terra. Le variabili esplicative prese in esame sono:

- `Pest`: quantità di un determinato pesticida con cui è stato trattato l'appezzamento
- `Colt`: la modalità di coltivazione della vite. La variabile è categoriale e ha due valori `Guyot` o `Alberello`
- `Pend`: la pendenza media dell'appezzamento
- `PggTot`: la pioggia accumulata nell'anno misurata da un pluviomerto posto nell'appezzamento

Vengono stimati i seguenti modelli:

Modello 1: $\text{resa} = \beta_0 + \beta_1(\text{Pest}) + \beta_2(\text{Colt}) + \beta_3(\text{Pend}) + \beta_4(\text{PggTot}) + \varepsilon$

Modello 2: $\text{resa} = \beta_0 + \beta_1(\text{Pest}) + \beta_2(\text{Colt}) + \beta_3(\text{Pend}) + \varepsilon$

Modello 3: $\text{resa} = \beta_0 + \beta_1(\text{Pest}) + \beta_2(\text{Colt}) + \beta_3(\text{PggTot}) + \varepsilon$

Modello 4: $\text{resa} = \beta_0 + \beta_1(\text{Pest}) + \beta_2(\text{Colt}) + \varepsilon$.

ε indica una variable casuale di media 0.

(i) Si indichi quali modelli sono annidati tra loro.

I modelli annidati fra loro sono

a. 1 e 2
b. 1 e 4
c. 2 e 4
d. 3 e 4

(ii) Si specifichi un nuovo modello 5 che abbia la caratteristica di essere annidato nel modello 1 (un solo modello a scelta). Si specifichi inoltre il sistema di verifica di ipotesi che si potrebbe utilizzare per verificare se la bontà di adattamento del Modello 5 sia significativamente diversa da quella del Modello 1.

Modello 5: resa = beta0 + beta1(Pest) + epsilon

Test ANOVA, in r: anova(modello_5, modello_1)

H0: beta2 = beta3= beta4 = 0 vs HA: any of (beta2 or beta3 or beta4) !=0

Si calcola la statistica F ed il relativo p-value

(iii) Si discuta brevemente che approccio si può utilizzare per confrontare la bontà di adattamento di modelli lineari quando questi non sono annidati tra loro.

Si possono utilizzare delle misure di bonta'che considerano la complessita' del modello come:

a. Adjusted R2, la versione pesata (sulla complessita' del modello) di R2
b. IC, AIC, BIC, che sono misure di bonta' di adattamento del modello basate sul calcolo della log verosimiglianza. BIC, penalizza molto i modelli poco parsimoniosi.
c. LOOCV RMSE
d. k-fold Cross-Validation,

Question 3 (7 points)

Un coltivatore d'uva studia l'effetto di un determinato trattamento sulla resa di due tipi di uve diverse. Diversi livelli del trattamento sono stati applicati a sei appezzamenti di terra e al momento della vendemmia viene misurata la resa. I dati sono qui riportati:

df

```
  resa tratt        uva
1 10.0   2.1    fragola
2 10.1   4.3    fragola
3  9.7   3.7    fragola
4  8.9   4.3 pizzutella
5 11.3   2.1 pizzutella
6 10.3   3.7 pizzutella
```

Il coltivatore utilizza i seguenti modelli statistici:

Modello 1: $\texttt{resa} = \beta_0 + \beta_1 \texttt{uva} + \varepsilon$

Modello 2: $\texttt{resa} = \beta_0 + \beta_1\ \texttt{uva} + \beta_2\ \texttt{tratt} + \varepsilon$

Modello 3: $\texttt{resa} = \beta_0 + \beta_1\ \texttt{uva} + \beta_2\ \texttt{tratt} + \beta_3\ \texttt{tratt} * \texttt{uva} + \varepsilon$

dove ε indica una variable casuale di media 0.

(i) Si scriva in forma esplicita la matrice di disegno del modello 2

nd<- data.frame(resa=c(10.0,10.1,9.7,8.9,11.3, 10.3), tratt=c(2.1, 4.3, 3.7,4.3, 2.1, 3.7), uva=c("fragola", "fragola", "fragola", "pizzutella", "pizzutella", "pizzutella"))

fit <- lm(resa~uva+tratt, data=nd); model.matrix(fit)

(Intercept)	uvapizzutella	tratt
1	0	2.1
1	0	4.3
1	0	3.7
1	1	4.3
1	1	2.1
1	1	3.7

(ii) Quale tra i tre modelli è più adatto a verificare se l'effetto del trattamento è diverso per i due tipi di uva? Si spieghi brevemente come è possibile fare questa verifica.

Il modello 3 e' il piu' adatto a verificare l'interazione tra il livello del trattamento ed il tipo di uva, poiche' la stessa e' inclusa all'interno del modello.

La verifica e' possibile tramite un test ANOVA dato che il modello 2 e' annidato nel modello 3.

(iii) Nei pannelli della Figura sottostante si creino delle visualizzazioni illustrative che mostrino schematicamente la relazione tra le variabili che è possibile modellare utilizzando i nei diversi modelli.